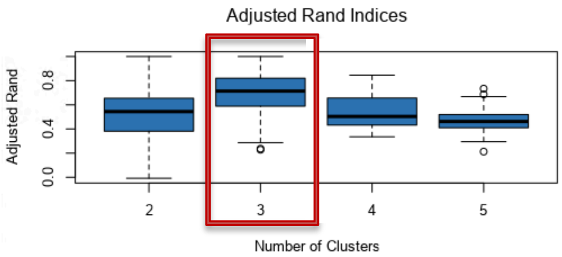
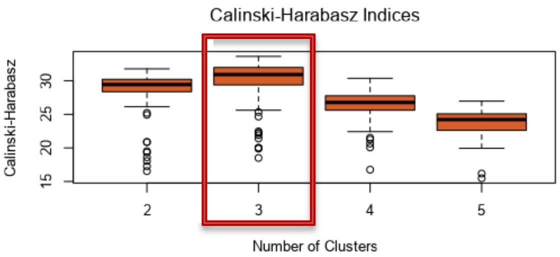
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

1. What is the optimal number of store formats? How did you arrive at that number?

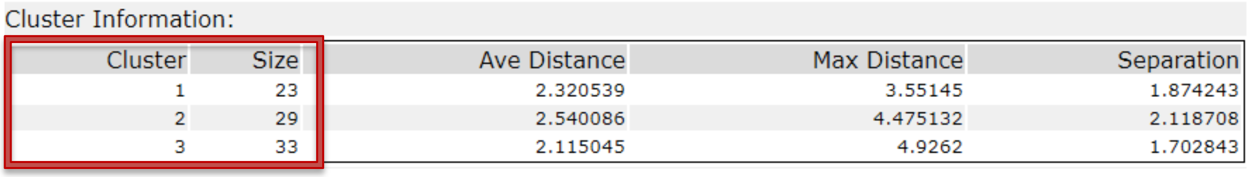
To determine optimal number of store cluster based on sales data I used **K-means clustering model.** According to ARI, 3 clusters show the highest coefficient of cluster stability as well as the highest CHI coefficient, which shows cluster compactness and distinctiveness. **Results suggest we should use 3 clusters.**





2. How many stores fall into each store format?

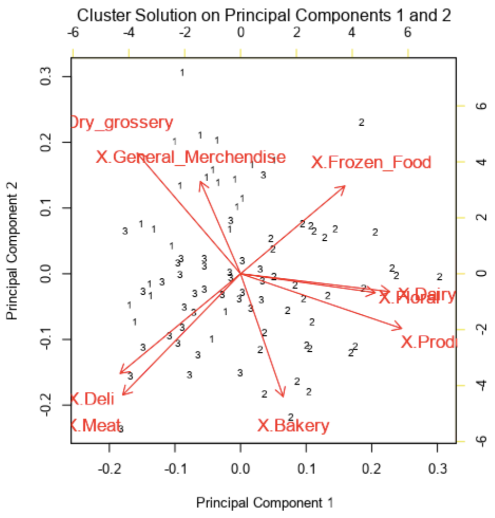
All stores fall into 3 clusters with size of 23, 29 and 33 stores accordingly.



Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

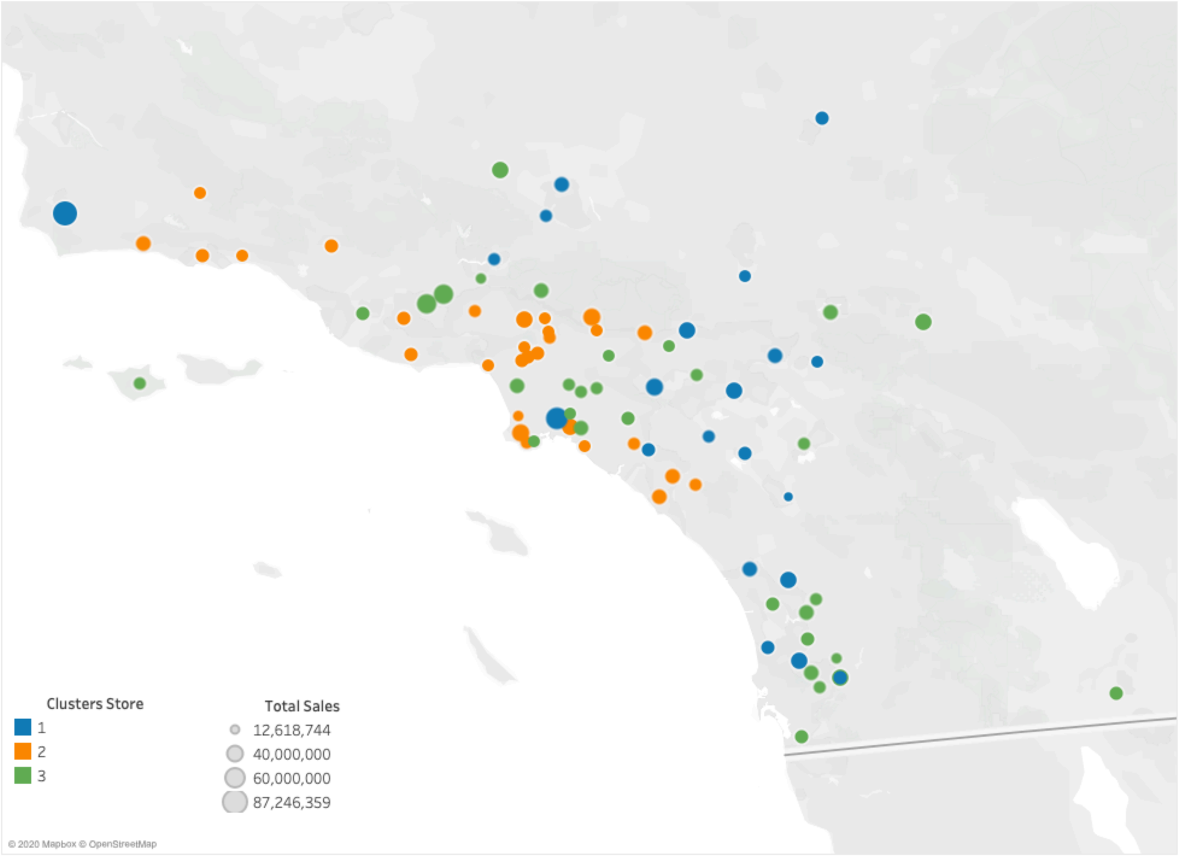


The Summary Report of the K-Means Clustering provide us with the Cluster solution on Principal Components plot. The plot illustrates the type of goods the clusters are oriented selling more. A **'cluster 1'** has highest total sales in 'General Merchandise' and 'Dry grocery' categories, while a **'cluster 3'** has a tendency to sell more 'Delicacy' and 'Meat' and a **'cluster 2'** the rest of categories.

4. **Please provide a Tableau visualization that shows the location of the stores, uses color to show cluster, and size to show total sales.**

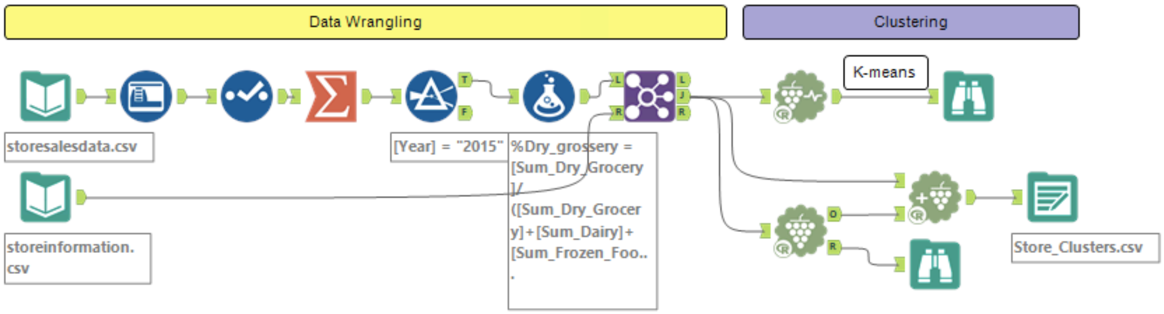
https://public.tableau.com/profile/lana1309#!/vizhome/StoreMap_15871527899620/Dashboard1?publish=yes

Store Map



Map based on Longitude (generated) and Latitude (generated). Color shows details about Clusters Store. Size shows sum of Total Sales. Details are shown for Zip.

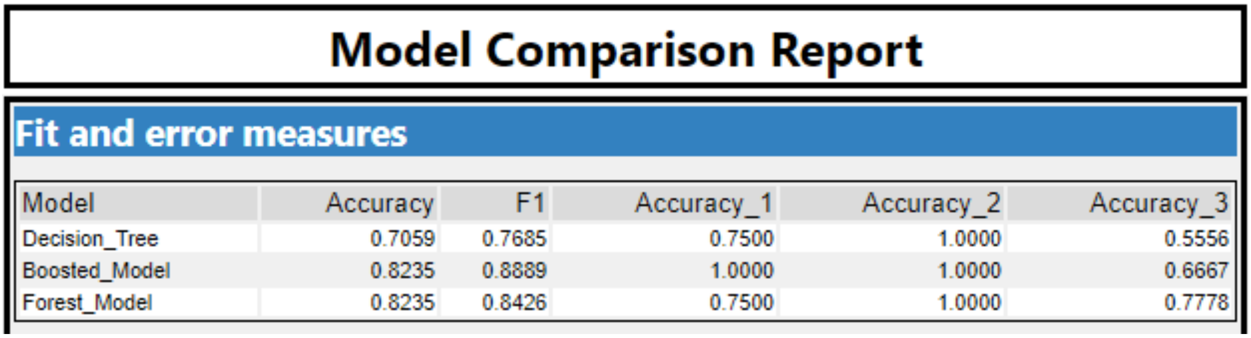
Alteryx workflow for clustering



Task 2: Formats for New Stores

1. **What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology?**

The results of the Model Comparison Report suggests we should use the **Boosted Model** for prediction as this model has the highest F1 coefficient (0.89).

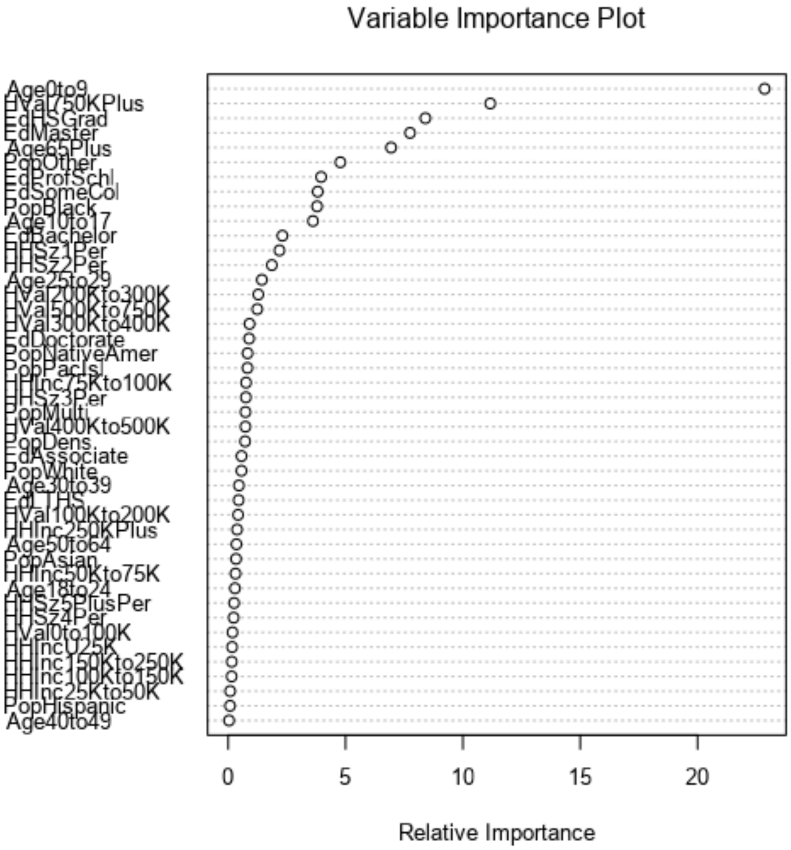


Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.7059	0.7685	0.7500	1.0000	0.5556
Boosted_Model	0.8235	0.8889	1.0000	1.0000	0.6667
Forest_Model	0.8235	0.8426	0.7500	1.0000	0.7778

2. **What are the three most important variables that help explain the relationship between demographic indicators and store formats? Please include a visualization.**

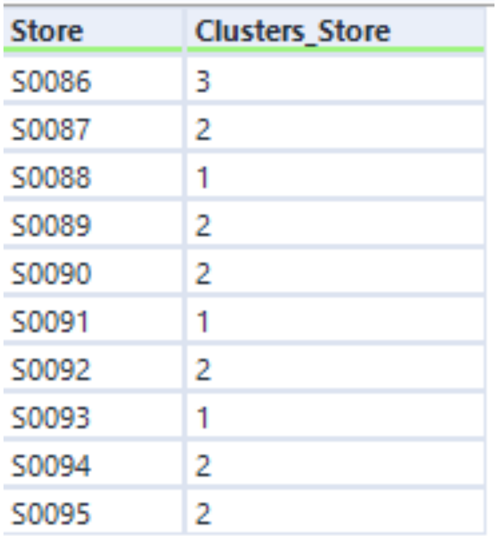


The Variable Importance Plot depicts:

- Presence of chilgren under 9 (‘Age 0 to 9’);
- Income higher 750K (‘HVal750KPlus’)
- High School Graduate (‘EdHSGrad’);

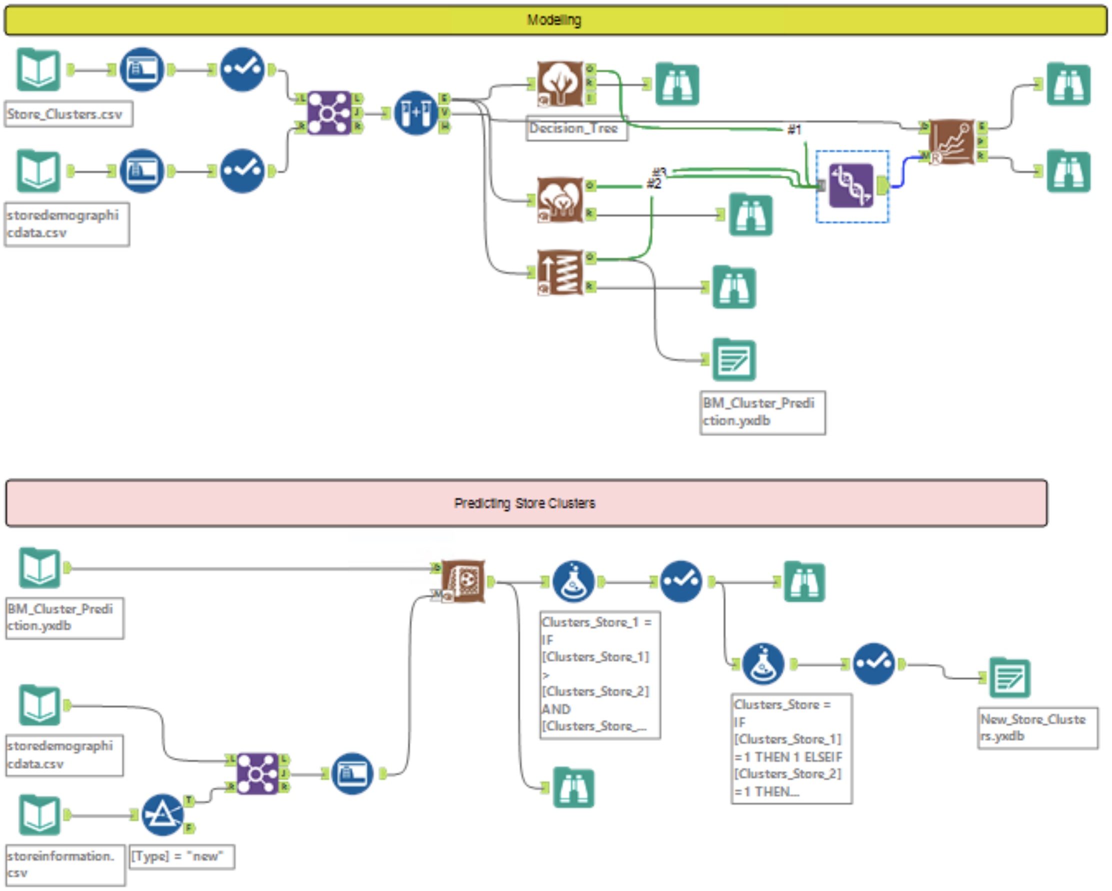
as the most important variables in Identification store clusters.

3. **What format do each of the 10 new stores fall into? Please fill in the table below.**



Store	Clusters_Store
S0086	3
S0087	2
S0088	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

Alteryx workflow for modeling and store clusters prediction


Modelling
Store_Clusters.csv
storedemographicdata.csv
Decision_Tree
#1
#3
#2
BM_Cluster_Prediction.yxdb
Predicting Store Clusters
BM_Cluster_Prediction.yxdb
storedemographicdata.csv
storeinformation.csv
[Type] = "new"
Clusters_Store_1 = IF [Clusters_Store_1] > [Clusters_Store_2] AND [Clusters_Store_...
Clusters_Store = IF [Clusters_Store_1] =1 THEN 1 ELSEIF [Clusters_Store_2] =1 THEN...
New_Store_Clusters.yxdb

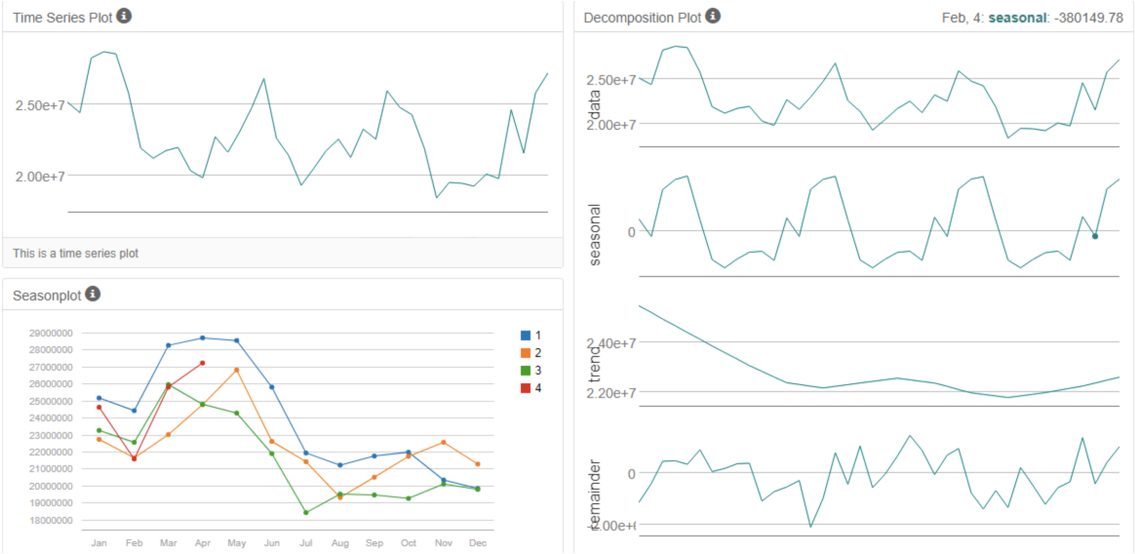
Task 3: Predicting Produce Sales

1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

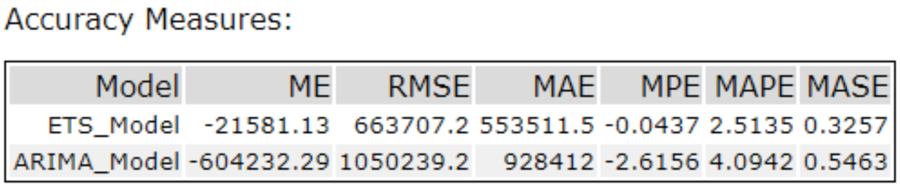
ETS model

The decomposition plot shows:

- **Error:** Significant fluctuations of over time, we should be applied it **multiplicatively (M)**
- **Trend:** Started as decreasing, then grows and drop over time, should be applied as **No** (**N**)
- **Seasonality:** Clearly seen on the graph with fluctuations of over time, should be applied it **multiplicatively (M)**



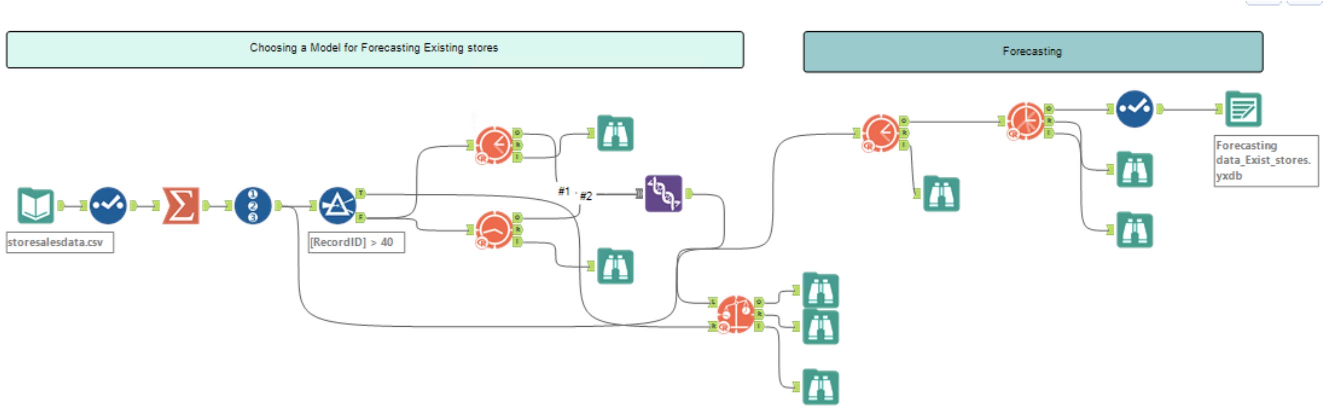
To forecast for the existing and new stores I chose between ETS(M,N,M) and ARIMA(1,0,0)(1,1,0)[12]. A report of the models comparison showed that ETS(M,N,M) model has lower coefficients of Accuracy Measures that means **ETS(M,N,M)** is more precise and should be used for forecasting for the existing stores.



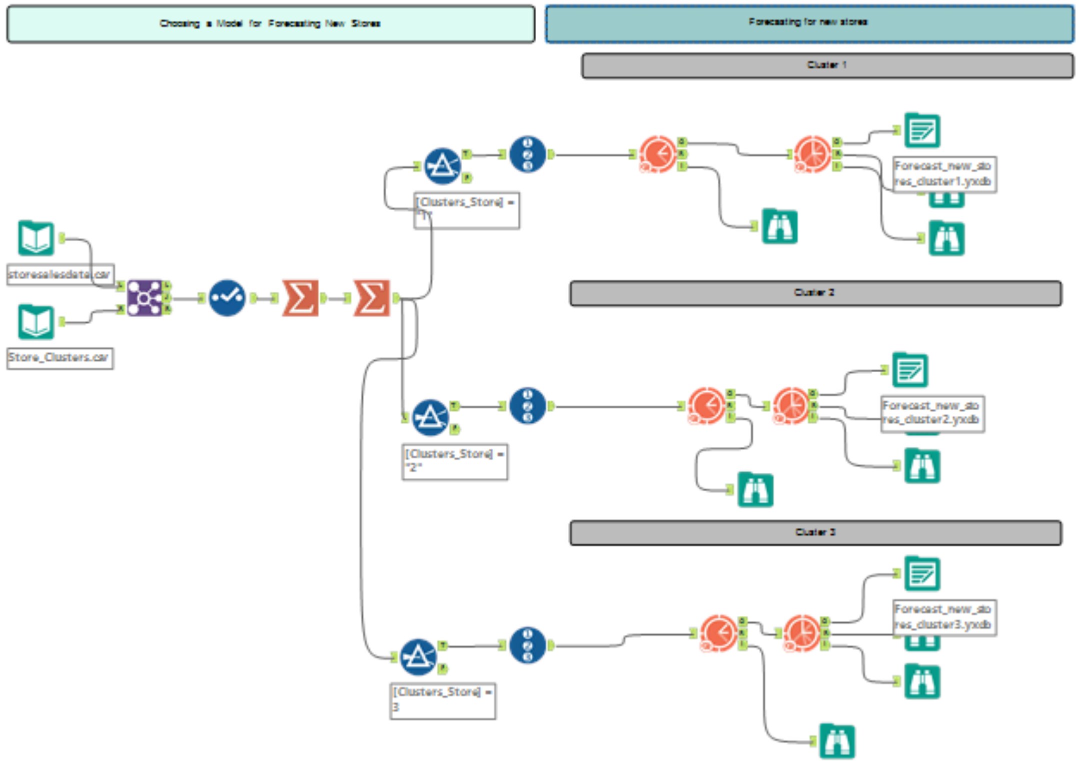
Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_Model	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
ARIMA_Model	-604232.29	1050239.2	928412	-2.6156	4.0942	0.5463

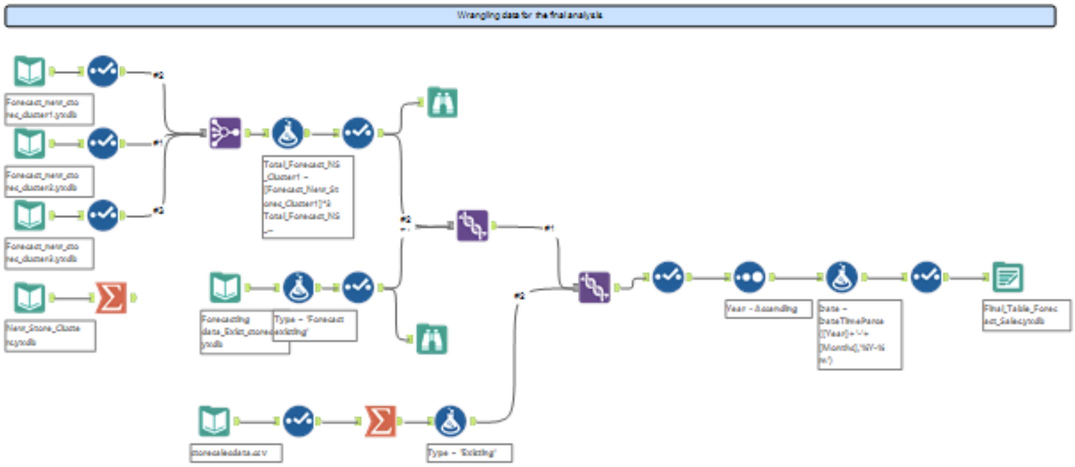
Workflow for forecasting for the existing stores


Choosing a Model for Forecasting Existing stores
Forecasting
storesalesdata.csv
[RecordID] > 40
#1
#2
Forecasting data_Exist_stores.yxdb

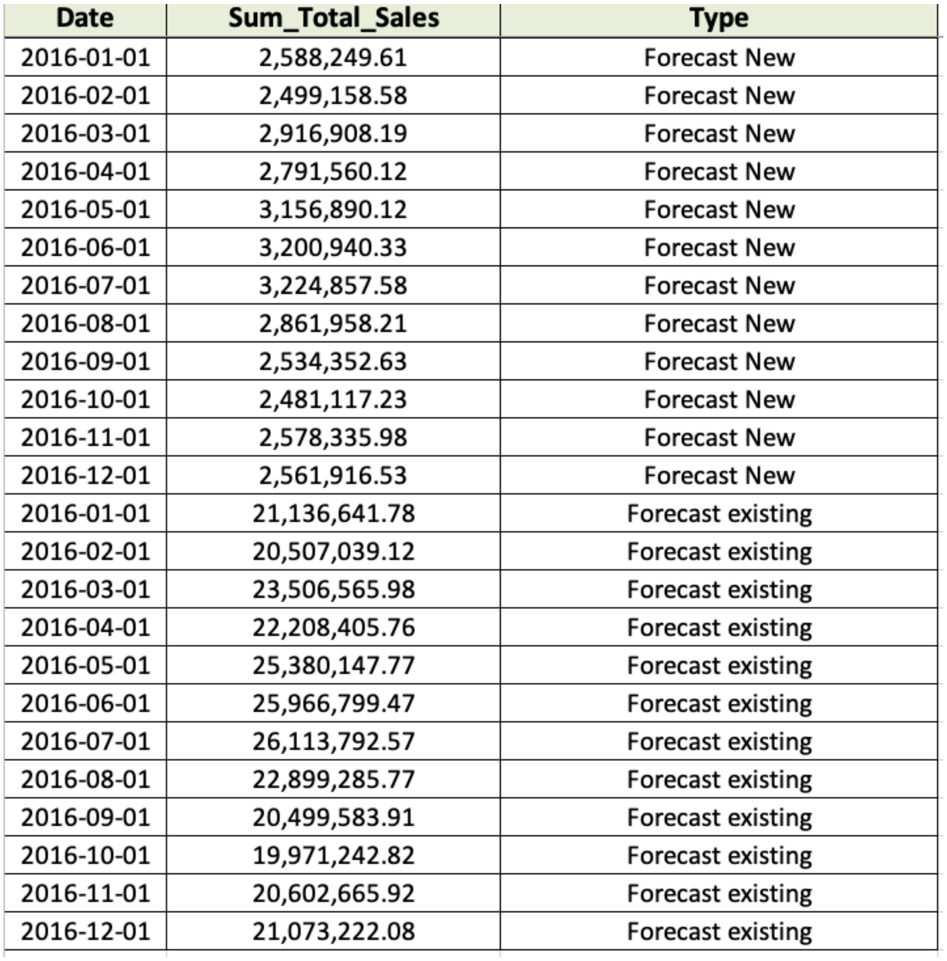
Workflow for forecasting for the new stores clusters


Choosing a Model for Forecasting New Stores
Forecasting for new stores
Cluster 1
Cluster 2
Cluster 3
storesalesdata.csv
Store_Clusters.csv
[Clusters_Store] = "1"
[Clusters_Store] = "2"
[Clusters_Store] = 3
Forecast_new_stores_cluster1.yxdb
Forecast_new_stores_cluster2.yxdb
Forecast_new_stores_cluster3.yxdb

Workflow for data wrangling for the final table “ Sales & forecast for existing and new stores”

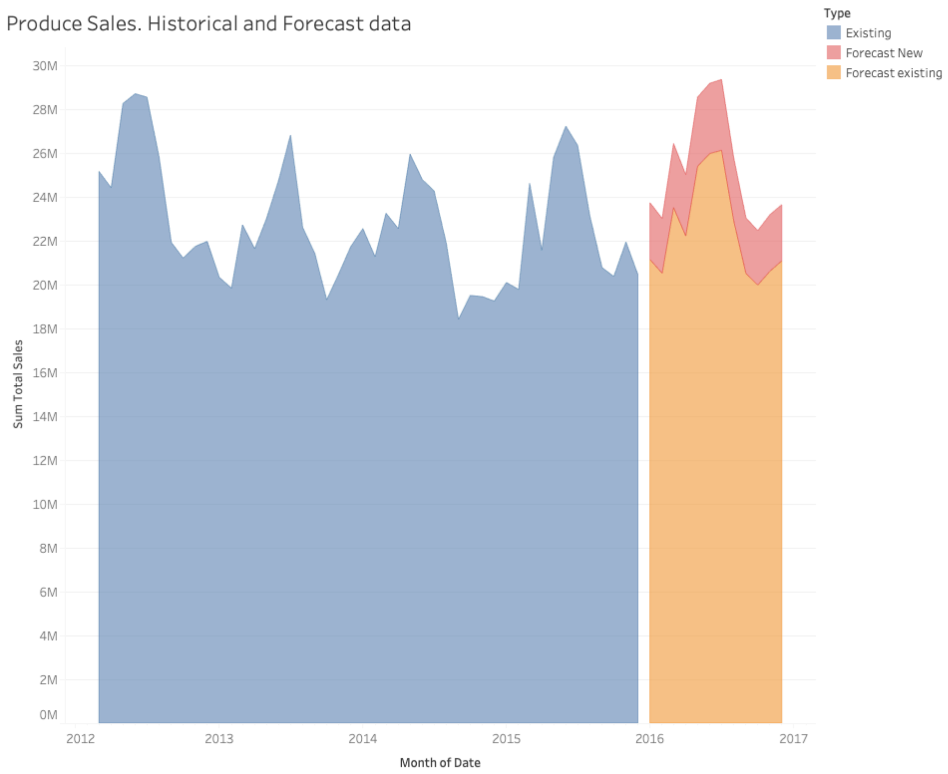

Wrangling data for the final analysis
Forecast_new_stores_cluster1.yxdb
Forecast_new_stores_cluster2.yxdb
Forecast_new_stores_cluster3.yxdb
New_Store_Clusters.yxdb
#2
#1
#3
Total_Forecast_NS_Cluster1 = [Forecast_New_Stores_Cluster1]*3 Total_Forecast_NS ...
Forecasting data_Exist_stores.yxdb
Type = 'Forecast existing'
#2
#1
#1
#2
storesalesdata.csv
Type = 'Existing'
Year - Ascending
Date = DateTimeParse([Year]+'-'+[Month],'%Y-%m')
Final_Table_Forecast_Sales.yxdb

2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.



Date	Sum_Total_Sales	Type
2016-01-01	2,588,249.61	Forecast New
2016-02-01	2,499,158.58	Forecast New
2016-03-01	2,916,908.19	Forecast New
2016-04-01	2,791,560.12	Forecast New
2016-05-01	3,156,890.12	Forecast New
2016-06-01	3,200,940.33	Forecast New
2016-07-01	3,224,857.58	Forecast New
2016-08-01	2,861,958.21	Forecast New
2016-09-01	2,534,352.63	Forecast New
2016-10-01	2,481,117.23	Forecast New
2016-11-01	2,578,335.98	Forecast New
2016-12-01	2,561,916.53	Forecast New
2016-01-01	21,136,641.78	Forecast existing
2016-02-01	20,507,039.12	Forecast existing
2016-03-01	23,506,565.98	Forecast existing
2016-04-01	22,208,405.76	Forecast existing
2016-05-01	25,380,147.77	Forecast existing
2016-06-01	25,966,799.47	Forecast existing
2016-07-01	26,113,792.57	Forecast existing
2016-08-01	22,899,285.77	Forecast existing
2016-09-01	20,499,583.91	Forecast existing
2016-10-01	19,971,242.82	Forecast existing
2016-11-01	20,602,665.92	Forecast existing
2016-12-01	21,073,222.08	Forecast existing

https://public.tableau.com/profile/lana1309#!/vizhome/Book1_15871531760470/Dashboard1?publish=yes


Produce Sales. Historical and Forecast data
Type
Existing
Forecast New
Forecast existing
Sum Total Sales
30M
28M
26M
24M
22M
20M
18M
16M
14M
12M
10M
8M
6M
4M
2M
0M
Month of Date